Protein splicing: occurrence, mechanisms and related phenomena

Yang Shao and Stephen BH Kent

An increasing number of proteins are thought to self-splice post-translationally on the level of the polypeptide, producing two separate proteins from one gene, neither of which is the protein predicted from the gene sequence. The recent elucidation of the mechanism of splicing has led to the identification of a number of post-translational protein modifications that use similar chemical pathways.

Address: Gryphon Sciences, 250 East Grand Avenue, Suite 90,
South San Francisco, CA 94080, USA.

Electronic identifier: **1074-5521-004-00187**

Chemistry & Biology March 1997, 4:187–194

Introduction

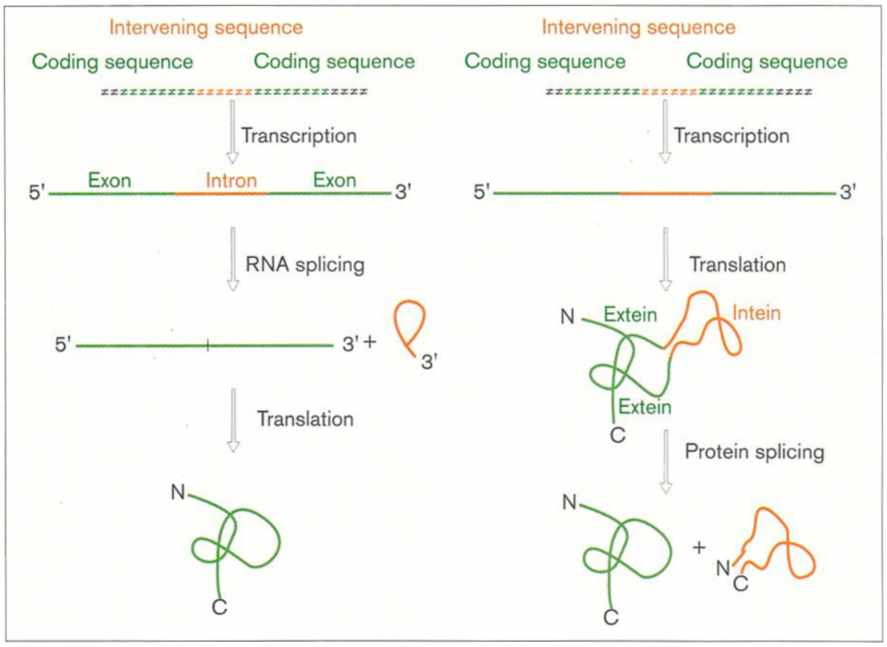
RNA splicing and protein splicing are two mechanisms by which the flow of information from a gene to its protein product can be modulated to yield a protein whose sequence is not strictly colinear with the gene. But, unlike RNA splicing which occurs at the level of an RNA precursor, protein splicing is a self-catalyzed process occurring on the level of the polypeptide in which an intervening polypeptide domain called an intein (int[ernal] [prot]ein) [1] is precisely excised from a precursor polypeptide. The flanking amino-terminal and carboxy-terminal polypeptide domains called exteins are concomitantly ligated together by the formation of a peptide bond to yield an active protein product (Fig. 1). As a radical post-translational processing event, protein splicing adds another layer of complexity to the mechanism of gene expression embodied in the 'central dogma' of molecular biology. Moreover, because a single gene gives rise to two separate protein products, protein splicing represents an exception to the idea that the final protein product is always colinear with the nucleotide sequence of the gene. Ever since its discovery as a new biological concept protein splicing has drawn worldwide attention and research on this topic has been advancing rapidly. This review is devoted to new results, new inteins and newly discovered related phenomena which were not included in the most recent review on this topic [2].

Occurrence

The first demonstrated example of protein splicing involved the protein product of the *Saccharomyces cerevisiae* TFP1 gene, which encodes the 69 kDa catalytic subunit of the vacuolar ATPase [3,4]. After that, further examples of protein splicing were discovered for proteins in all three kingdoms of life: in eukaryotes, the 69 kDa subunit of vacuolar ATPase of the yeast *Candida tropicalis* [5]; in prokaryotes, the RecA proteins of the mycobacteria *Mycobacterium tuberculosis* [6,7] and *Mycobacterium leprae* [8], and in Archaea, the DNA polymerases of the extremely thermophilic archaebacteria *Thermococcus litoralis* [9] and *Pyrococcus* species GB-D [10].

The fact that the exteins found in these examples of protein splicing share no obvious homology and can be replaced by other polypeptides, indicates that they are not involved in the catalytic events of protein splicing. On the other hand, the inteins, which range from 350 to 550 amino acid residues, show moderate amino acid sequence similarity. Moreover, a comparison of the amino acid residues at the junctions between the intein and the flanking exteins reveals a striking conservation of amino acid residues. Just downstream (i.e. towards the carboxyl terminus of the

Figure 1



A comparison between RNA splicing and protein splicing. In the mechanism of RNA splicing, shown on the left, the transcription of the coding sequence immediately precedes the RNA splicing. RNA splicing thus occurs at the level of the RNA precursor and the intron is excised before translation of the RNA to the protein. In contrast, protein splicing, shown on the right, occurs at the level of the protein. In protein splicing, a polypeptide domain called an intein is excised from the middle region of a precursor polypeptide to give two proteins that are not colinear with the coding sequence.

polypeptide chain) of each splice junction, there is always an amino acid residue with a hydroxyl or sulfhydryl sidechain, such as serine, threonine or cysteine. The carboxyl terminus of the inteins consists of the sequence His–Asn preceded by several hydrophobic residues (Table 1). In addition to these four highly conserved positions, these inteins contain the two dodecapeptide motifs characteristic of homing endonuclease [11]. The homing endonuclease activity of the mature inteins, released by protein splicing, is thought to provide the inteins with gene mobility [12].

Based on these conserved amino acids and motifs, more open reading frames containing intein-like intervening sequences have been found in searches of genomic databases. For example, the *pps1* open reading frame of *M. leprae* [11]; the GyrA protein in *M. leprae*, *Mycobacterium kansasii*, *Mycobacterium flavescens*, *Mycobacterium gordonae* [13] and *Mycobacterium xenopi* (GenBank accession number U67876); the DNA polymerases of the extremely thermophilic archaebacterium *Pyrococcus* species KOD1 (Genbank accession number D29671); the *cclp* gene in *Chlamydomonas eugametos* chloroplast [14]; The dnaB helicase in *Porphyra purpurea* [15] and *Synechocystis* [16]; and, most recently, the 18 inteins in the genome of *Methanococcus jannaschii* [17] (Table 1). Interestingly, among these putative inteins some have Gly–Asn instead of the usual His–Asn at their carboxyl terminus (Table 1, numbers 3,29) while others lack the dodecapeptide motifs characteristic of homing endonuclease, making their sizes unusually small (Table 1, numbers 4,13). As new genome sequences are disclosed, we expect to find more inteins.

Mechanisms

Since the discovery of protein splicing, several mechanisms have been proposed. In the absence of structural information about the inteins on which to base the proposals, these mechanisms were based on the types of reactions that might be possible for the conserved amino acid residues at the splice junctions. Several putative chemical pathways were proposed to be involved in the process of protein splicing. Wallace [18] suggested that both splice junctions underwent N–O/N–S acyl rearrangements, followed by nucleophilic attack of the downstream α-amino group on the upstream ester or thioester, then hydrolysis of the downstream ester or thioester. Cooper *et al.* [19] proposed that protein splicing starts with cleavage between the intein and the C-extein by cyclization of asparagine at the downstream splice junction followed by a transpeptidation reaction between the upstream splice junction and the amino terminus of the C-extein. Clarke [20] proposed that protein splicing begins with the attack of the downstream asparagine at the upstream splice junction to form a succinimide, however. Xu *et al.* [21] proposed two possible mechanisms for protein splicing. One starts with a serine proteinase-like attack of the downstream serine at the upstream splice junction, followed by the downstream asparagine cyclization and O–N acyl rearrangement. The other mechanism, shown in Figure 2, involves four steps. Step 1: formation of an ester intermediate by an N–O acyl

Table 1

Known and putative inteins.

Number	Protein	Organism	Amino terminus	Carboxyl terminus	Number of residues	Reference
1	VMA intein	*S. cerevisiae*	C	HN/C	454	[3,4]
2	VMA intein	*C. tropicalis*	C	HN/C	471	[5]
3	*clpP* intein	*C. eugametos*	C	GN/S	456	[14]
4	*dnaB* intein	*P. purpurea*	C	HN/S	150	[15]
5	*dnaB* intein	*Synechocystis*	C	HN/S	429	[16]
6	*recA* intein	*M. tuberculosis*	C	HN/C	440	[6]
7	*recA* intein	*M. leprae*	C	HN/S	365	[7]
8	*pps1* intein-1	*M. leprae*	C	HN/S	386	[11]
9	*gyrA* intein	*M. leprae*	C	HN/T	420	[13]
10	*gyrA* intein	*M. kansasii*	C	HN/T	420	[13]
11	*gyrA* intein	*M. flavescens*	C	HN/T	421	[13]
12	*gyrA* intein	*M. gordonae*	C	HN/T	420	[13]
13	*gyrA* intein	*M. xenopi*	C	HN/T	198	U67876*
14	*pol* intein-1	*T. litoralis*	S	HN/S	538	[9]
15	*pol* intein-2	*T. litoralis*	S	HN/T	390	[9]
16	*pol* intein-1	*Pyrococcus* sp. GB-D	S	HN/S	537	[10]
17	*pol* intein-3	*Pyrococcus* sp. KOD	S	HN/S	536	D29671*
18	*pol* intein-2	*Pyrococcus* sp. KOD	C	HN/S	360	D29671*
19	*pol* intein-1	*M. jannaschii*	C	HN/S	369	[17]
20	*pol* intein-2	*M. jannaschii*	S	HN/S	476	[17]
21	*hyp* intein-1	*M. jannaschii*	C	HN/C	392	[17]
22	*hyp* intein-2	*M. jannaschii*	C	HN/C	488	[17]
23	IF intein-2	*M. jannaschii*	C	HN/T	546	[17]
24	TFIIB intein	*M. jannaschii*	S	HN/T	335	[17]
25	Pep Syn intein	*M. jannaschii*	C	FN/C	412	[17]
26	RNR intein-1	*M. jannaschii*	S	HN/T	453	[17]
27	RNR intein-2	*M. jannaschii*	S	HN/T	533	[17]
28	Rpol A″ intein	*M. jannaschii*	S	HN/T	471	[17]
29	Rpol A′ intein	*M. jannaschii*	C	GN/C	452	[17]
30	UDP GD intein	*M. jannaschii*	C	HN/C	454	[17]
31	Helicase intein	*M. jannaschii*	C	HN/S	501	[17]
32	GF-6P intein	*M. jannaschii*	C	HN/S	499	[17]
33	*r-gyr* intein	*M. jannaschii*	C	HN/C	494	[17]
34	RFC intein-1	*M. jannaschii*	C	HN/T	548	[17]
35	RFC intein-2	*M. jannaschii*	S	HN/S	436	[17]
36	RFC intein-3	*M. jannaschii*	C	HN/C	543	[17]

*GenBank accession number.

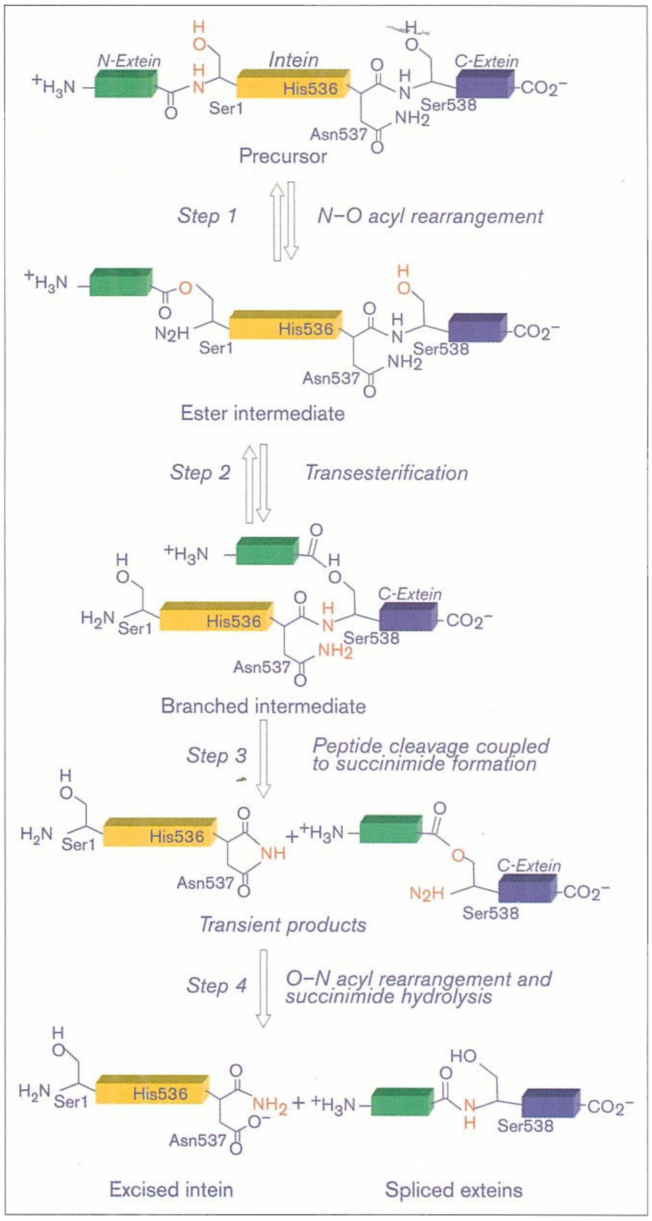
rearrangement of the conserved serine residue at the upstream splice junction. Step 2: formation of a branched intermediate by transesterification involving attack by the hydroxyl sidechain of Ser1 at the downstream splice junction on the ester formed in Step 1. Step 3: excision of the intein by peptide bond cleavage coupled to succinimide formation involving the conserved asparagine residue at the downstream splice junction. Step 4: spontaneous O–N acyl rearrangement of the transitory ligation product from an ester to the more stable amide.

Xu and coworkers [22] at New England Biolabs Inc. ingeniously developed an *in vitro* protein splicing system by inserting the intein from the extreme thermophile *Pyrococcus* sp. GB-D between two foreign proteins, the maltose binding protein and *Dirofilaria immitis* paramyosin, to give a three-part chimera. The fusion protein was then expressed at low temperature, providing a source of unspliced precursor. This breakthrough not only provided indisputable evidence that protein splicing is post-translational and self-catalytic, but also made it possible to intercept the steps in the process of protein splicing and to isolate and characterize key intermediates. The mechanism shown in Figure 2 is the most consistent with all the reported experimental data obtained from this *in vitro* splicing system, and will be discussed in detail.

Step 1: formation of an ester intermediate by an N–O acyl rearrangement

A unique feature of this mechanism is the occurrence of a linear ester intermediate in which the N-extein domain is esterified at the carboxyl terminus with the sidechain hydroxyl of Ser1 of the intein. This results from an N–O acyl rearrangement at the upstream splice junction and is consistent with the fact that Ser1 is conserved as the first amino acid in some inteins. The N–O acyl rearrangement, first described for aliphatic amino alcohols by Bergmann *et al.* [23], has been observed in peptides adjacent to serine

Figure 2



Mechanism of protein splicing (adapted from [21,30,36]). The mechanism shown involves four steps: the formation of an ester intermediate, the formation of a branched intermediate, excision of the intein, and spontaneous O–N acyl rearrangement to form the more stable amide forms of the excised intein and spliced exteins. See text for further details.

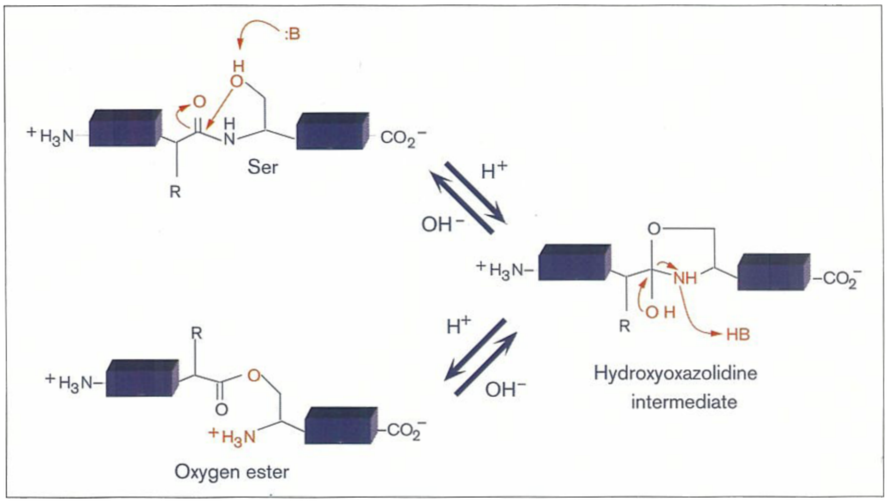
and threonine residues under strongly acidic conditions [24,25]. The rearrangement is thought to involve a hydroxyoxazolidine intermediate [25,26] and is promoted by acid, which protonates the free amino group, thus shifting the equilibrium to the ester form. At neutral or alkaline pH, however, the N–O acyl rearrangement is rapidly reversed because the equilibrium favors the amide rather than the ester form (Fig. 3). Such reactions can therefore occur in nature only when the ester immediately undergoes a second reaction that displaces the unfavorable equilibrium, as in the biosynthesis of pyruvoyl-dependent enzymes such as histidine decarboxylase from *Lactobacillus*, where the amino-terminal pyruvoyl moiety is generated by an N–O acyl rearrangement at an internal serine residue followed by a β-elimination step [27].

Shao *et al.* [28] used the strategy of replacing Ser1 of the intein with cysteine to explore the involvement of an N–O acyl rearrangement in protein splicing. Cysteine is also conserved as the first amino acid in some inteins (Table 1), and its thioesters are much more susceptible to attack by nitrogen nucleophiles at neutral pH than oxygen esters [29]. If an ester intermediate involving the sidechain of Ser1 were involved in the splicing process, replacement of the serine hydroxyl by a cysteine thiol group should significantly enhance the rate of protein cleavage at the upstream splice junction by hydroxylamine at neutral pH. Shao *et al.* [28] observed that cleavage of the unspliced precursor protein, in which the amino terminal cysteine had been replaced with cysteine, could be enhanced by nitrogen nucleophiles such as hydroxylamine and ethylene diamine. They isolated and characterized the hydroxamate produced from the hydroxylaminolysis at the upstream splice junction. This observation suggested that the peptide bond at the upstream splice junction undergoes an N–O or N–S acyl rearrangement to yield either an ester or a thioester. The rearrangement of the upstream N–O acyl is also supported by the mutagenesis data of Xu and Perler [30].

Step 2: formation of a branched intermediate by transesterification

The branched intermediate is formed by transesterification when the hydroxyl sidechain of Ser1 of the C-extein at the downstream splice junction attacks the ester formed in Step 1. Step 2 resembles the action of a serine proteinase on an ester substrate where the branched intermediate resembles the acyl-enzyme intermediate. When serine proteinases such as chymotrypsin and trypsin hydrolyze ester and thioester substrates, the rates of hydrolysis of the ester and thioester substrates are about 1000 times slower than that for the amide substrates and the rate-limiting step is the hydrolysis of the acyl-enzyme intermediate [31]. If this is extrapolated to protein splicing, it would suggest that the branched intermediate would accumulate in the process of protein splicing. A unique feature of the postulated branched intermediate is that it has two amino termini. Using their *in vitro* protein splicing system, Xu *et al.* [22] successfully isolated a transient band by sodium dodecyl sulfate polyacrylamide gel electrophoresis. Amino-terminal amino acid sequence analysis of this band found two amino acids in each cycle of the analysis, suggesting that it has two amino termini; this is consistent with the characteristics of the postulated branched intermediate.

Figure 3



Mechanism of N–O acyl rearrangement. Acidic conditions promote the formation of the hydroxyoxazolidine intermediate and therefore the ester formation. Neutral or alkaline pH reverses the rearrangement because the equilibrium favours the amide rather than the ester form.

The branched intermediate was the first intermediate of protein splicing to be identified, and its discovery laid the cornerstone for further mechanistic elucidation.

Step 3: excision of the intein

Protein cleavage by succinimide formation has been found in the aging of the α_A subunit of bovine α-crystallin [32]. Step 3 displaces the unfavorable equilibrium between the ester and amide form, present in Step 1 of protein splicing, making the whole splicing process irreversible. In support of this, Shao *et al.* [33] synthesized a peptide with a carboxy-terminal aminosuccinimide residue, corresponding to the putative carboxyl terminus of the excised intein derived from the thermostable DNA polymerase of *Pyrococcus* sp. GB-D. A methionine residue was then inserted a few amino acids upstream towards the carboxyl terminus of the intein. After the *in vitro* splicing, the intein was cleaved by cyanogen bromide. The synthetic aminosuccinimide peptide was compared with the carboxy-terminal cyanogen bromide peptide of the excised intein and found to be indistinguishable in terms of its chromatographic properties, high resolution mass spectrum, and colorimetric assay involving its reaction with hydroxylamine. This definitively established that protein splicing is accompanied by the cyclization of asparagine to yield an aminosuccinimide residue at the carboxyl terminus of the excised intein and that this unusual residue is therefore a natural constituent of the excised intein.

Step 4: spontaneous O–N acyl rearrangement

In Step 4, the transitory ligation products undergo spontaneous O–N acyl rearrangements to form the more stable amides. The equilibrium of N–O acyl rearrangements favors the amide at neutral or high pH. For this reason, there are no known examples of O–N acyl rearrangements in proteins other than those thought to occur in the course of protein splicing. Examples of O–N and S–N acyl rearrangements are seen in the chemical synthesis of proteins, however [34,35]. Model studies showed that under splicing conditions the O–N acyl rearrangement occurs with a half life of 7 min, much faster than *in vitro* protein splicing [22].

After elucidating the mechanism of protein splicing involving thermophilic inteins, Chong *et al.* [36] used a similar strategy and discovered that the intein of the mesophilic yeast VMA also adopted the same chemical pathway in the process of splicing. In this case, because there are conserved cysteines at both splicing junctions, N–S acyl rearrangements are involved rather than N–O acyl shifts.

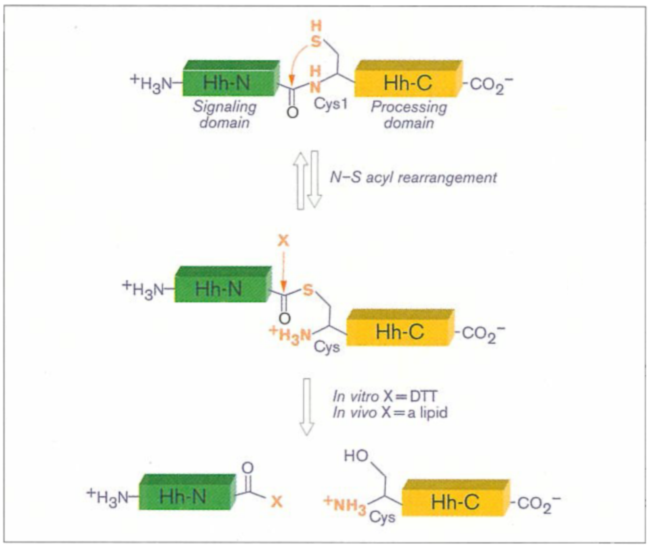
Related phenomena

Since the elucidation of the mechanisms of protein splicing, there have been reports of post-translational modifications of proteins involving similar chemical pathways. The most recent of these reports is the autoprocessing of Sonic Hedgehog proteins.

Sonic Hedgehog proteins are responsible for the patterning of a variety of embryonic structures in vertebrates and invertebrates. The *Drosophila* hedgehog gene has been shown to generate two predominant protein species by a self-catalyzed internal cleavage of a larger precursor. Autocatalytic processing mediated by the carboxy-terminal domain of the Hedgehog protein precursor (Hh-C) generates an amino-terminal domain product (Hh-N) that accounts for all known signaling activity [37,38]. The cleavage site of Hedgehog protein is immediately before a cysteine and conserved amino acid sequences are found at the cleavage site and at the upstream splice junction of two yeast inteins [39]. It is therefore likely that an N–S acyl rearrangement, analogous to that involved in protein splicing, occurs in the maturation of the Sonic Hedgehog protein. Experimental evidence [40,41] has shown that the autoprocessing reaction proceeds via an internal thioester intermediate formed through N–S acyl rearrangement and nucleophilic attack by the hydroxyl group of a cholesterol molecule. It has been shown that cholesterol is the lipophilic moiety that is covalently attached to the amino-terminal signaling domain during autoprocessing and that the carboxy-terminal domain acts as an intramolecular cholesterol transferase [41]. This cholesterol molecule increases the hydrophobic character of the signaling domain, influencing its spatial and subcellular distribution [40] (Fig. 4).

Another phenomenon which adopts similar chemical pathways to protein splicing is the activation of glycosylasparaginase from *Flavobacterium meningosepticum* [42]. The peptide bond between Asp151 and Thr152 undergoes an N–O acyl rearrangement and the resulting ester intermediate is rapidly hydrolyzed. This self-catalyzed cleavage to produce two subunits is obligatory for the function of glycosylasparaginase; the newly formed amino-terminal threonine of the β-subunit is essential for enzyme activity. Additional evidence for this mechanism was provided by the crystal structure of its human homolog [43]. The structure of the mature human protein suggests that it would be difficult for any external proteolytic enzymes to reach the threonine in the deep and narrow pocket of the funnel-like active site.

Figure 4

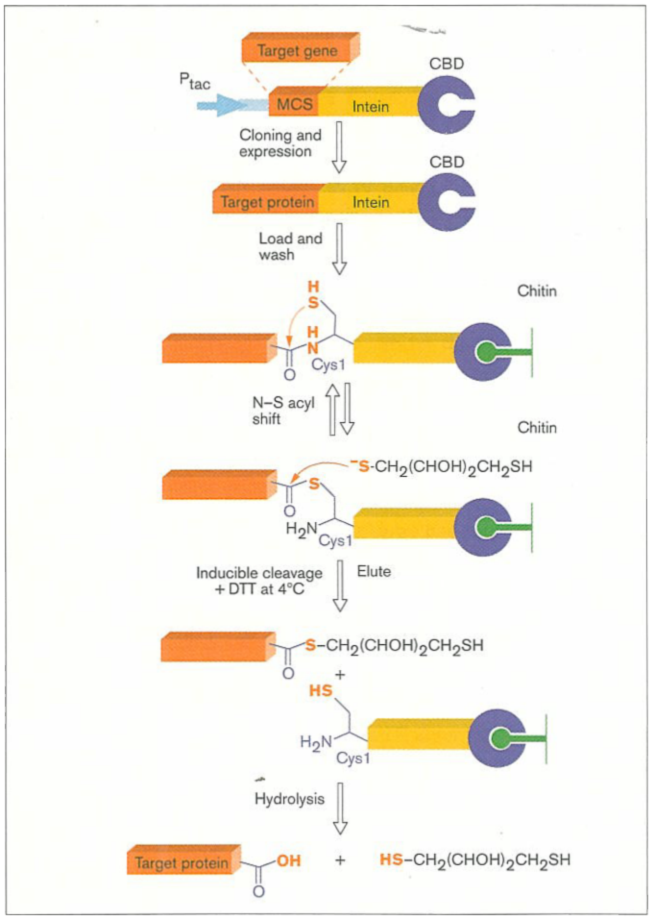


Mechanism of Hedgehog protein autoprocessing (adapted from [41]). Autocatalytic processing mediated by the carboxy-terminal domain of the Hedgehog protein precursor (Hh-C) generates an amino-terminal domain product (Hh-N). Because the cleavage site is immediately before a cysteine and conserved amino acid sequences are found at the cleavage site and at the upstream splice junction of two yeast inteins, it is likely that an N–S acyl rearrangement, analogous to that involved in protein splicing, occurs in the maturation of the Hedgehog protein.

Glycosylasparaginase belongs to the newly recognized structural superfamily of amino-terminal nucleophile amidohydrolases, which includes penicillin acylase [44], the proteasome β-subunit [45], and glutamine 5-phosphoribosyl-1-pyrophosphate amidotransferase [46]. These three enzymes are all amidotransferases. Penicillin acylase cleaves the amide bond between the variable group and the β-lactam ring in penicillins to produce 6-aminopenicillanic acid. Like glycosylasparaginase, penicillin acylase is an αβ dimer formed by the cleavage of a single polypeptide chain precursor [43]. Glutamine 5-phosphoribosyl-1-pyrophosphate amidotransferase is the regulatory enzyme of the *de novo* synthesis of purine nucleotides. It is a homotetramer, with each subunit composed of two folding domains [45]. The proteasome is the central enzyme of non-lysosomal protein degradation. It contains 14 copies each of the structurally similar α- and β-subunits [44]. Not only do these enzymes have three-dimensional structures similar to glycosylasparaginase, but they also have the unusual feature of using an amino-terminal nucleophile (serine, threonine or cysteine) as their catalytic center [47]. This suggests that similar self-catalyzed N–O and N–S acyl rearrangements leading to protein cleavage may also occur in the maturation or activation of these enzymes.

The elucidation of the mechanism of protein splicing inspired the scientists at New England Biolabs Inc. to invent an ingenious purification kit called IMPACT I™ for recombinant proteins. This kit uses the intein from the *S. cerevisiae* VMA1 gene and works according to the mechanism shown in Figure 5. The intein has been modified so that at low temperatures in the presence of thiols (such as dithiothreitol (DTT), β-mercaptoethanol or cysteine) it undergoes a self-cleavage reaction at its amino terminus. The gene encoding the target protein is inserted into a multiple cloning site (MCS) of a pCYB vector to create a fusion between the carboxyl terminus of the target gene and the amino terminus of the gene encoding the intein. The DNA encoding a small 5 kDa chitin binding domain (CBD) from *Bacillus circulans* has been added to the carboxyl terminus of the intein for affinity purification of the expressed fusion protein. Expression of the three-part chimeric fusion protein is controlled by an IPTG-inducible P_{tac} promotor. When crude extracts of cells from an inducible *Escherichia coli* expression system are passed through a chitin column, the fusion protein binds to the chitin column while all other contaminants are washed through the column. The fusion protein is then induced to undergo an intein-catalyzed self-cleavage on the column by overnight incubation at 4°C in the presence of DTT or β-mercaptoethanol. The target protein is released into

Figure 5



Schematic illustration of the IMPACT I™ system (adapted from IMPACT I™ manual, New England Biolabs Inc.). The gene encoding the target protein is inserted into a multiple cloning site (MCS) of a pCYB vector. DNA encoding a small chitin binding domain (CBD) is attached to the carboxyl terminus of the intein. The three-part fusion protein is expressed and then the crude cell extracts are passed through a chitin column. The fusion protein binds to the chitin column while all the contaminants are washed through the column. The fusion protein is then induced to undergo an intein-catalyzed self-cleavage on the column thereby releasing the target protein with the eluant and leaving the intein–chitin binding domain bound to the column.

the eluant while the intein–chitin binding domain fusion remains bound to the column. Thus, native proteins can be purified without the use of proteinases and the carboxyl terminus of the target protein can be labeled with, for example, isotopes or fluorescent tags.

Perspectives

Although the biochemical mechanism of protein splicing has been determined (Fig. 2), there is currently no structural information on the protein splicing precursor. It still remains to be determined what the catalytic groups are and how they catalyze each partial reaction. Questions, such as whether only the neighboring groups in the primary sequence catalyze the N–O and N–S acyl rearrangements or whether the catalytic machinery is composed of amino acids at distant positions of the polypeptide chain as in the serine proteinases, still remain unanswered. Because the intervening sequences are multifunctional proteins that also function as homing endonucleases, amino acid sequence alignments cannot provide any information towards answering these questions. Efforts are being made by scientists to crystallize the protein splicing precursor, in the hope that the crystal structure may shed light on the catalytic machinery for protein splicing.

With the discovery of more and more putative inteins from searches of the genomic database, we need to ask whether in these cases the intervening sequences are in fact spliced out of their polypeptide precursors. For those inteins containing a Gly–Asn sequence at their carboxyl terminus (Table 1, #3,29), it would also be interesting to determine whether cofactors are needed to compensate for the lack of the histidine that is normally conserved at this position.

What is the biological function of inteins? Inteins are probably homing endonucleases because, with the exception of the dnaB intein from *P. purpurea* and the gyrA intein from *M. xenopi* (Table 1, #4,13), they all contain the LAGLI-DADG motif. This motif is characteristic of homing endonucleases encoded by introns of the rRNA genes in archaea and bacteria, and by group I introns in eukaryotes. With their homing activities, the inteins and their coding sequence may represent the most primitive parasite, which does not even encode an entire protein but relies on the host for the initiation and termination of protein synthesis. To compensate for this extreme form of genetic economy, the intein may have evolved the catalytic potential to excise itself from host proteins, thereby assuring the survival of the host, even if the selfish DNA has insinuated itself into an essential gene. Thus, like introns, inteins can be considered manifestations of selfish DNA that may survive by being able to invade and multiply without significantly compromising their host. Studies of the inteins lacking the homing endonuclease activities, such as the dnaB intein from *P. purpurea* and gyrA intein from *M. xenopi* (Table 1, #4,13), might tell us more about the origin and function of inteins and whether this radical post-translational protein modification may be more widespread in nature.

Acknowledgements

The authors wish to thank Dr Perler at New England Biolabs for her critical reading of this manuscript and for her insightful comments and suggestions.

References

1. Perler, F.B., *et al.*, & Belfort, M. (1994). Protein splicing elements: inteins and exteins - a definition of terms and recommended nomenclature. *Nucleic Acids Res.* **22**, 1125–1127.
2. Cooper, A.A. & Stevens, T.H. (1995). Protein splicing: self-splicing of genetically mobile elements at the protein level. *Trends Biochem. Sci.* **20**, 351–356.

3. Kane, P.M., *et al.*, & Stevens, T.H. (1990). Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651–657.
4. Hirata, R., *et al.*, & Anraku, Y. (1990). Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)- translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae. J. Biol. Chem.* **265**, 6726–6733.
5. Gu, H.H., Xu, J., Gallagher, M. & Dean, G.E. (1993). Peptide splicing in the vacuolar ATPase subunit A from *Candida tropicalis. J. Biol. Chem.* **268**, 7372–7381.
6. Davis, E.O., Sedgwick, S.G. & Colston, M.J. (1991). Novel structure of the recA locus of *Mycobacterium tuberculosis* implies processing of the gene product. *J. Bacteriol.* **173**, 5653–5662.
7. Davis, E.O., Jenner, P.J., Brooks, P.C., Colston, M.J. & Sedgwick, S.G. (1992). Protein splicing in the maturation of *M. tuberculosis* recA protein: a mechanism for tolerating a novel class of intervening sequence. *Cell* **71**, 201–210.
8. Davis, E.O., Thangaraj, H.S., Brooks, P.C. & Colston, M.J. (1994). Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. *EMBO J.* **13**, 699–703.
9. Hodges, R.A., Perler, F.B., Noren, C.J. & Jack, W.E. (1992). Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucleic Acids Res.* **20**, 6153–6157.
10. Perler, F.B., *et al.*, & Jannasch. J. (1992). Intervening sequences in an Archaea DNA polymerase gene. *Proc. Natl Acad. Sci. USA* **89**, 5577–5581.
11. Pietrokovski, S. (1994). Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci.* **3**, 2340–2350.
12. Gimble, F.S. & Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. *Nature* **357**, 301–306.
13. Fsihi, H., Vincent, V. & Cole, S.T. (1996). Homing events in the gyrA gene of some mycobacteria. *Proc. Natl Acad. Sci. USA* **93**, 3410–3415.
14. Huang, C., *et al.*, & Liu, X.-Q. (1994). The *Chlamydomonas* chloroplast clpP gene contains translated large insertion sequences and is essential for cell growth. *Mol. Gen. Genet.* **244**, 151–159.
15. Reith, M.E. & Munholland, J. (1995). Complete nucleotide sequence of of the *Porphyra purpurea* chloroplast genome. *Plant Mol. Biol. Rep.* **13**, 333–335.
16. Pietrokovski, S. (1996). A new intein in cyanobacteria and its significance for the spread of inteins. *Trends Genet.*, **12**, 287–288.
17. Bult, C.J., *et al.*, & Venter, J.C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273**, 1058–1073.
18. Wallace, C.J. (1993). The curious case of protein splicing: mechanistic insights suggested by protein semisynthesis. *Protein Sci.* **2**, 697–705.
19. Cooper, A.A., Chen, Y.-J., Lindorfer, M.A. & Stevens, T.H. (1993). Protein splicing of the yeast TFP1 intervening protein sequence: a model for self-excision. *EMBO J.* **12**, 2575–2583.
20. Clarke, N.D. (1994). A proposed mechanism for the self-splicing of proteins. *Proc. Natl Acad. Sci. USA* **91**, 11084–11088.
21. Xu, M.-Q., *et al.*, & Perler, F.B. (1994). Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. *EMBO J.* **13**, 5517–5522.
22. Xu, M.-Q., Southworth, M.W., Mersha, F.B., Hornstra, L.J., & Perler, F.B. (1993). *In vitro* protein splicing of purified precursor and the identification of a branched intermediate. *Cell* **75**, 1371–1377.
23. Bergmann, M., Brand, E. & Dreyer, F. (1921). Synthase von alpha, beta-diglyceriden und symmetrischen triglyceriden. [Title translation: Synthesis of alpha, beta-diglyceride and unsymmetric triglyceride.] *Berichete* **54**, 936–965.
24. Desnuelle, P. & Casal, A. (1948). Sur la moindre resistance a l'hydrolyse acide des liason peptidiques situees à côté d'une fonction hydroxyle. [Title translation: On the least resistance to acidic hydrolysis of peptide bonds situated next to a hydroxyl group.] *Biochim. Biophys. Acta* **2**, 64–75.
25. Elliott, D.F. (1952). A search for specific chemical methods for fission of peptide bonds. *Biochem. J.* **50**, 542–550.
26. Iwai, K. & Ando, T. (1967). N-O acyl shifts. *Meth. Enzymol.* **11**, 262–282.
27. Van Poelje, P.D. & Snell, E.E. (1990). Cloning, sequencing, expression, and site-directed mutagenesis of the gene from *Clostridium perfringens* encoding pyruvoyl-dependent histidine decarboxylase. *Biochemistry* **29**, 132–139.
28. Shao, Y., Xu, M.-Q. & Paulus, H. (1996). Protein splicing: evidence for an N-O acyl rearrangement as the initial step in the splicing process. *Biochemistry* **35**, 3810–3815.
29. Jencks, W.P., Cordes, S. & Carriulo, J. (1960). The free energy of thioester hydrolysis. *J. Biol. Chem.* **235**, 3608–3614.
30. Xu, M.-Q. & Perler, F.B. (1996). The mechanism of protein splicing and its modulation by mutation. *EMBO J.* **15**, 5146–5153.
31. Fersht, A. (1985). In *Enzyme Structure and Mechanism* (2nd edn), pp. 201–208 and 405–413. W.H. Freeman and Company, NY, USA.
32. Voorter, C.E.M., De Haard-Hoekman, W.A., Van den Oetelaar, P.J.M., Bloemendal, H. & De Jong, W.W. (1988). Spontaneous peptide bond cleavage in aging alpha-crystallin through a succinimide intermediate. *J. Biol. Chem.* **263**, 19020–19023.
33. Shao, Y., Xu, M.-Q. & Paulus, H. (1995). Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. *Biochemistry* **34**, 10844–10850.
34. Liu, C.F. & Tam, J.P. (1994). Peptide segment ligation strategy without use of protecting groups. *Proc. Natl Acad. Sci. USA* **91**, 6584–6588.
35. Dawson, P.E., Muir, T.W., Clark-Lewis, I. & Kent, S.B. (1994). Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779.
36. Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F.B. & Xu, M.-Q. (1996). Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an *in vitro* splicing system. *J. Biol. Chem.* **271**, 22159–22168.
37. Lee, J.J., Ekker, S.C., Kessler, D.P., Porter, J.A., Sun, B.I. & Beachy, P.A. (1994). Autoproteolysis in Hedgehog protein biogenesis. *Science* **266**, 1528–1537.
38. Porter, J.A., *et al.*, & Beachy, P.A. (1995). The product of Hedgehog autoproteolytic cleavage active in local and long-range signalling. *Nature* **374**, 363–366.
39. Koonin, E.V. (1995). A protein splice-junction motif in Hedgehog family proteins. *Trends Biochem. Sci.* **20**, 141–142.
40. Porter, J.A., *et al.*, & Beachy, P.A. (1996). Hedgehog patterning activity: role of a lipophilic modification mediated by the carboxy-terminal autoprocessing domain. *Cell* **86**, 21–34.
41. Porter, J.A., Young, K.E. & Beachy, P.A. (1996). Cholesterol modification of Hedgehog signaling proteins in animal development. *Science* **274**, 255–259.
42. Guan, C., *et al.*, & Comb, D. (1996). Activation of glycosylasparaginase. Formation of active N-terminal threonine by intramolecular autoproteolysis. *J. Biol. Chem.* **271**, 1732–1737.
43. Oinonen, C., Tikkanen, R., Rouvinen, J. & Peltonen, L. (1995). Crystal structure of human glycosylasparaginase. *Nat. Struct. Biol.* **2**, 1102–1106.
44. Duggleby, H.J., Tolley, S.P., Hill, C.P., Dodson, E.J., Dodson, G. & Moody, P.C. (1995). Penicillin acylase has a single-amino acid catalytic centre. *Nature* **373**, 264–268.
45. Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. & Huber, R. (1995). Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 3.4 Å resolution. *Science* **268**, 533–539.
46. Smith, J.L., *et al.*, & Satow, Y. (1994). Structure of the allosteric regulatory enzyme of purine biosynthesis. *Science* **264**, 1427–1433.
47. Artymiuk, P.J. (1995). A sting in the (N-terminal) tail. *Nat. Struct. Biol.* **2**, 1035–1038.